Perceptual learning for speech in noise after application of binary time-frequency masks

Mahnaz Ahmadi[a)]
Department of Psychology, Utah State University, 2810 Old Main Hill, Logan, Utah 84322-2810

Vauna L. Gross
Department of Communicative Disorders and Deaf Education, Utah State University, 1000 Old Main Hill, Logan, Utah 84322-1000

Donal G. Sinex[b)]
Department of Psychology, Utah State University, 2810 Old Main Hill, Logan, Utah 84322-2810

(Received 8 February 2012; revised 4 November 2012; accepted 14 January 2013)

Ideal time-frequency (TF) masks can reject noise and improve the recognition of speech-noise mixtures. An ideal TF mask is constructed with prior knowledge of the target speech signal. The intelligibility of a processed speech-noise mixture depends upon the threshold criterion used to define the TF mask. The study reported here assessed the effect of training on the recognition of speech in noise after processing by ideal TF masks that did not restore perfect speech intelligibility. Two groups of listeners with normal hearing listened to speech-noise mixtures processed by TF masks calculated with different threshold criteria. For each group, a threshold criterion that initially produced word recognition scores between 0.56–0.69 was chosen for training. Listeners practiced with one set of TF-masked sentences until their word recognition performance approached asymptote. Perceptual learning was quantified by comparing word-recognition scores in the first and last training sessions. Word recognition scores improved with practice for all listeners with the greatest improvement observed for the same materials used in training.
 [http://dx.doi.org/10.1121/1.4789896]

PACS number(s): 43.71.Gv, 43.66.Dc, 43.72.Dv [CYE] Pages: 1687–1692

I. INTRODUCTION

Listeners with hearing loss (HL) report difficulty understanding speech in the presence of competing sounds; in fact, this is often said to be their most common complaint. In laboratory studies, listeners with HL require a higher speech-to-noise ratio (SNR) than listeners with normal hearing to achieve the same speech-recognition performance (Plomp and Mimpen, 1979; Turner, 2006; Helfer and Freyman, 2008). The development of more-effective signal-processing methods for reducing the impact of noise on intelligibility is an important scientific and clinical goal.

One promising method for removing noise and improving speech intelligibility involves the calculation and application of a binary time-frequency (TF) mask (Anzalone *et al*., 2006; Brungart *et al*., 2006; Li and Loizou, 2008; Wang *et al*., 2008, 2009; Kjems *et al*., 2009; Narayanan and Wang, 2010). Generation of an "ideal" TF mask requires independent knowledge of the speech to be enhanced, hence the technique is currently not suitable for applications outside the laboratory such as noise reduction for hearing aids. However, techniques based on TF masks that do not require prior knowledge of the target signal may eventually be developed to the point at which translational applications to audiology are feasible; Wang (2008) has provided an extensive review of the potential uses of TF masks in hearing-aid design.

In most studies of noise reduction by TF masks, mask parameters are varied, and the effectiveness of the mask is evaluated by measuring speech recognition for a small number of trials or sessions. Although some combinations of TF mask parameters restore perfect intelligibility for SNRs as unfavorable as −60 dB (Kjems *et al*., 2009), recognition scores may decrease all the way to zero for some other combinations. The local threshold criterion used to define the mask is especially important. The local criterion and its use in constructing a TF mask are explained in detail in Sec. II B. Low recognition scores may arise because a high threshold removes speech information along with noise; alternatively, low scores may occur because the mask will fail to remove enough noise if the criterion is too low (Brungart *et al*., 2006). In either case, but especially in the first, the signal presented for identification may have an abnormal quality that interferes with recognition. Training has been found to increase understanding of speech when intelligibility has been compromised by factors including noise or hearing loss (Sweetow and Palmer, 2005; Burk *et al*., 2006; Burk and Humes, 2007, 2009; Choi *et al*., 2009; Tyler *et al*., 2010). The goal of the present experiment was to determine whether the intelligibility of TF-masked speech would improve with practice for conditions for which initial intelligibility was low.

[a)]Present address: Division of Communication and Auditory Neuroscience, House Research Institute, 2100 West Third Street, Los Angeles, CA 90057.
[b)]Author to whom correspondence should be addressed. Present address: Department of Communication Disorders, University of Canterbury, Christchurch 8140, New Zealand. Electronic mail: donal.sinex@canterbury.ac.nz

II. METHODS

A. Stimuli

Speech stimuli consisted of sentences from the coordinate response measure (CRM) corpus (Bolia *et al.*, 2000). Each sentence includes three key words and has the form "Ready (call sign) go to (color) (number) now." CRM sentences are well-suited for a study of practice effects since the same sentence can be presented repeatedly without regard for possible effects of context, set size, or variation in difficulty. Sentences in the complete CRM corpus include eight call signs, four colors, and eight numbers, each of which is spoken by 4 male and 4 female talkers. The experiments reported here used a subset of 32 sentences produced by a single male talker (number 0 in the original corpus) and using the call sign "Baron." The sentences were mixed with noise whose spectrum matched the long-term spectrum of all CRM sentences produced by male talkers.

B. Model parameters

Ideal TF masks were generated with procedures similar to those described by Brungart *et al.* (2006), Kjems *et al.* (2009), and Wang *et al.* (2008, 2009) The ideal TF mask is a matrix of cells, each defined by a particular time interval ("frame") and frequency band. The stimulus within each cell is evaluated and judged to be dominated either by speech or by noise; that information is used to process the speech-noise mixture to preserve speech and reject noise. The frame length was 20 msec, and successive frames overlapped by 10 msec. The frequency bands were 0.20 octave wide, covered the range from 0.1 to 6.0 kHz with 16 bands per octave, and were implemented with second-order Butterworth filters; with these specifications the filters had bandwidths that are comparable to those of auditory filters (Moore and Glasberg, 1987).

The masks were generated with SNR fixed at −12 dB. Figure 1 shows spectrograms calculated for one representative CRM sentence, in quiet [Fig. 1(a)] and mixed with noise [Fig. 1(b)]. With few exceptions, the noise completely obscures the spectral and temporal features of the sentence in the figure; perceptually, word recognition scores for CRM sentences presented at this SNR with no additional processing have been reported to be less than 0.1 (Eddins and Li, 2012). To generate a TF mask, the root-mean-square (RMS) amplitude of the speech signal in each time-frequency cell was compared to the RMS amplitude of the masking noise in the same cell. The difference in levels was tested against a criterion SNR; in this study, the threshold criterion was the only mask parameter that varied across conditions. If the SNR in the cell was equal to or greater than the criterion, the cell was assumed to be dominated by speech and the gain for that cell was set to 1; otherwise, the cell was assumed to be dominated by noise and the gain was set to 0. The criterion SNR can be described in absolute terms, that is, as the actual SNR that must be exceeded to count the cell as dominated by speech. This SNR has been referred to as the "local criterion" (LC) by Brungart *et al.* (2006). The same criterion can be normalized with respect to the overall SNR of the mixture; in

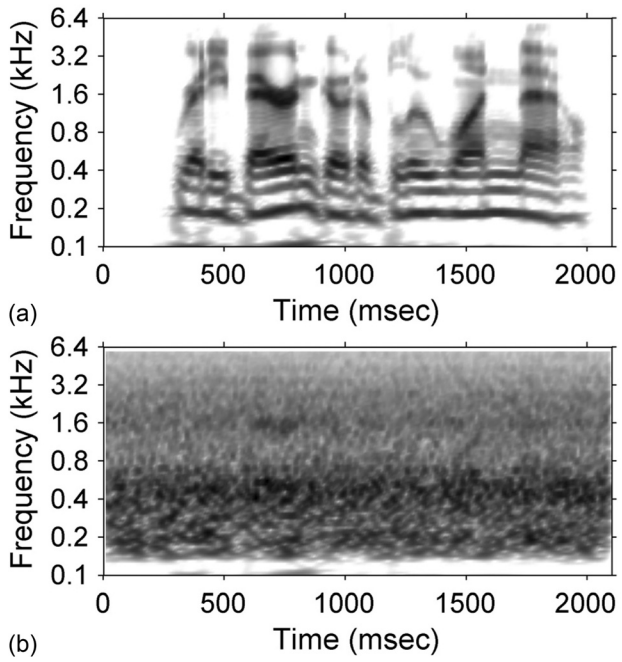

FIG. 1. Spectrograms of a representative CRM sentence in (a) quiet and (b) mixed with noise at −12 dB SNR.

that case, the criterion is referred to as a "relative criterion" (RC; Kjems *et al.*, 2009). The two values are related to one another by the simple equation RC = LC − SNR. When recognition scores are shown as a function of RC, identification is largely independent of overall SNR (Kjems *et al.*, 2009). For any particular SNR, recognition varies with RC (Brungart *et al.*, 2006; Kjems *et al.*, 2009; Li and Loizou, 2008). Intelligibility is essentially perfect for RC in the range from approximately −20 to 0 dB but decreases for lower and higher values of RC.

In the experiment reported here, the threshold criterion was always specified as a RC. Ideal TF masks calculated for three values of RC are shown in Fig. 2. The TF mask in Fig. 2(a) was generated with RC = −10 dB, a value that typically produces nearly perfect identification (Brungart *et al.*, 2006; Li and Loizou, 2008). Consistent with that, the pattern of cells in the TF mask for which gain was set to 1 strongly resembles the spectrogram for the sentence in quiet [Fig. 1(a)]. The TF mask shown in Fig. 2(b) was calculated with RC = +10 dB. In this case, the density of cells set to have a gain of 1 was much lower. As noted above, identification scores decrease for values of RC in this region because the TF mask rejects too much of the speech signal; this low-density pattern is consistent with that. Figure 2(c) shows a TF mask calculated with RC = −30 dB. The density of cells set to have a gain of 1 was much higher in this example; that is also consistent with the previous statement that values of RC in this range generally fail to reject enough of the noise masker.

The TF mask was then applied to the speech-noise mixture to generate the waveforms that were presented for identification. To apply the mask, the speech-noise mixture was analyzed in the same matrix of time-frequency cells as before. For each cell, if the gain of the mask was 1, the

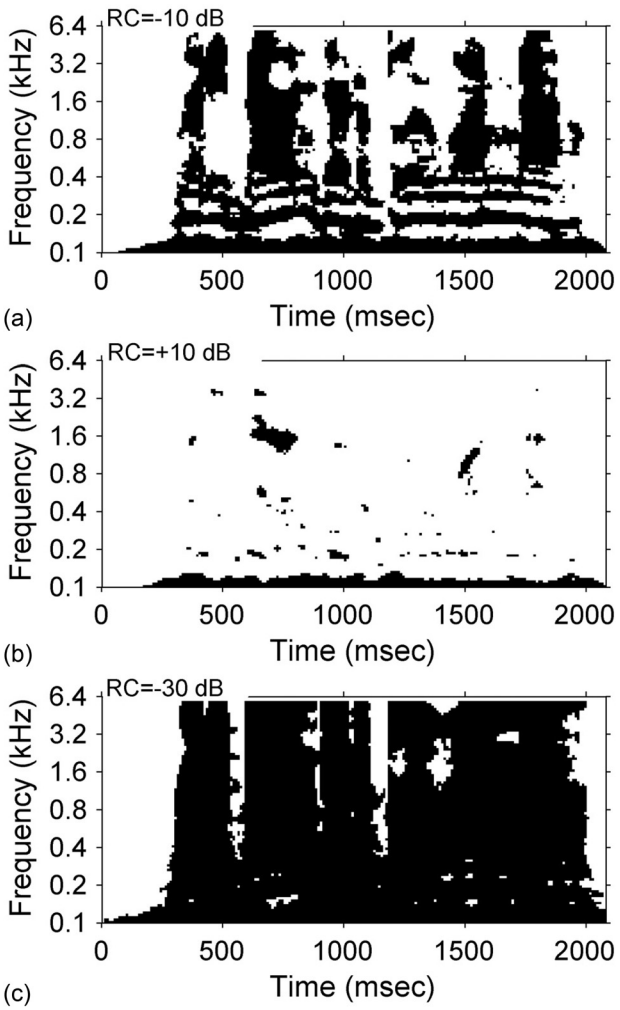

FIG. 2. TF masks calculated from the speech-noise mixture shown in Fig. 1. Cells for which gain was set to 1 are shown in black. (a) RC = −10 dB, (b) RC = +10 dB, (c) RC = −30 dB.

waveform in that band and time frame was added to the output waveform. If the gain of the mask was 0, the waveform in that cell was discarded. The output waveform that resulted after all cells had been processed was presented for identification.

C. Listeners

Data were obtained from 8 listeners (2 females and 6 males) between the ages of 20 and 32 years (mean 23.37 years). None reported any history of hearing difficulties, and all had thresholds below 20 dB SPL (sound pressure level) for pure tones at octave frequencies from 0.25–4.0 kHz. The listeners provided informed consent and were paid an hourly wage.

D. Psychophysical procedure

Psychophysical measurements were made in a single-walled sound-treated booth (Industrial Acoustics Corporation, Bronx, NY). Sentences from the CRM corpus were mixed with noise and processed with a TF mask as described previously. The waveforms were sent from a high-quality sound card (Gina 3 G, Echo Digital Audio, Santa Barbara, CA) to Sennheiser HD 280 (Sennheiser Electronic Corp., Old Lyme, CT) pro circumaural headphones. Waveforms were always presented diotically. After each sentence presentation, the listeners identified the spoken color and number combination by using the computer mouse to select the appropriate colored digit on the computer screen. Feedback was provided after each trial. Sentences were presented in blocks of 32 trials; each block included one repetition of each color-number combination.

All listeners completed a pre-training phase, a training phase, and a post-training phase. During the pre-training phase, TF masks were generated with selected values of RC that were thought likely to produce less-than-perfect word recognition. Listeners completed a single block of 32 trials at each of 6–10 values of RC in the range from −35 to +25 dB. For the training phase, RC for the TF mask was fixed at one of two levels for which word recognition scores obtained in the pre-training phase fell in the range from approximately 0.56–0.69. Group 1 heard sentences processed with a TF mask generated with RC = +10 dB. Group 2 differed from Group 1 only in that the TF masks were generated with RC = −30 dB. During the training sessions, listeners completed an average of 30 blocks (960 trials) per day for 6 to 11 days within a 3 to 4 week period. The post-training phase was similar to the pre-training phase in that sentences processed with TF masks generated with a range of RC values were presented.

A trial was scored as correct if the listener accurately reported both the color and the number. The proportion of words scored as correct in a block or blocks of trials is referred to as the word recognition score.

All procedures were reviewed and approved by the Institutional Review Board at Utah State University.

III. RESULTS

Word recognition scores obtained during the pre-training phase are shown in Fig. 3. The overall dependence of word recognition scores on RC was consistent with the pattern reported by others (Brungart *et al.*, 2006; Li and Loizou, 2008). Word recognition was essentially perfect for values of RC between −20 and −10 dB, but was lower for the other tested values. Values of RC between −10 dB and +10 dB were omitted since as noted above the purpose of the pre-training phase was to identify values of RC where

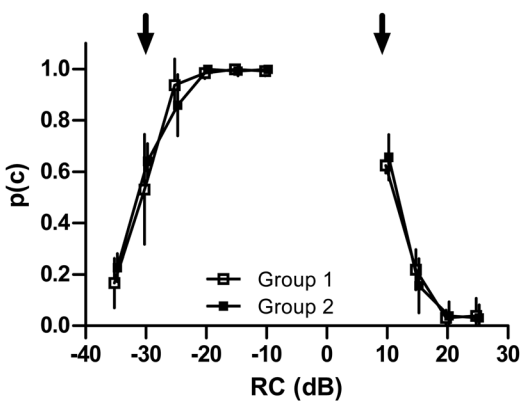

FIG. 3. Recognition of TF-masked CRM sentences before training. Open symbols: mean of 4 listeners in Group 1. Filled symbols: mean of 4 listeners in Group 2. Error bars are standard deviations across listeners. The arrows identify the two values of RC that were selected for the training phase.

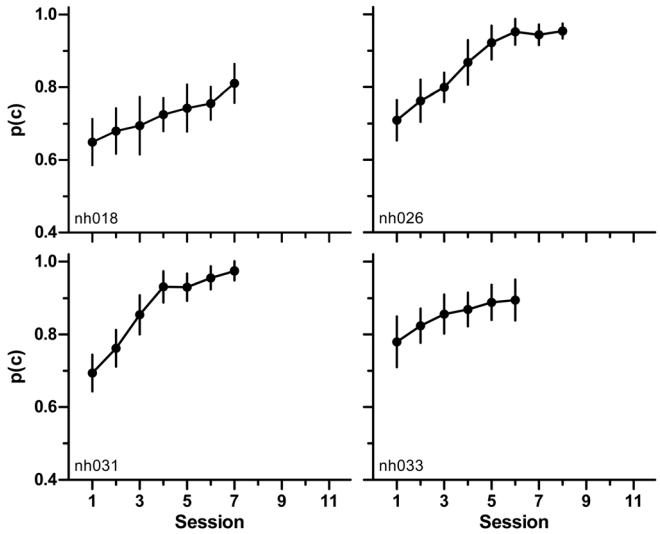

FIG. 4. Recognition of CRM sentences processed by TF masks during the training phase. Each panel presents data for 1 of the 4 listeners in Group 1, for whom TF masks were generated with RC = +10 dB. Symbols show the mean word recognition score calculated across all the blocks in a daily session. Error bars are standard deviations.

performance improvements could potentially occur. As expected, there were no differences between the two groups during pre-training. The arrows in Fig. 3 mark the two values of RC that were selected and used to process sentences during the subsequent training phase.

Figures 4 and 5 depict intelligibility scores for the individual listeners during the training phase. Data from Group 1, trained with sentences processed by TF masks with RC = +10 dB, are shown in Fig. 4. Data from Group 2, trained with sentences processed by TF masks with RC = −30 dB, are shown in Fig. 5. For all listeners, identification scores for TF-masked sentences increased during training. However, individual listeners improved at different rates and required different amounts of practice to approach asymptotic performance.

Learning effects were evaluated by comparing intelligibility scores obtained on the first and last days of training.

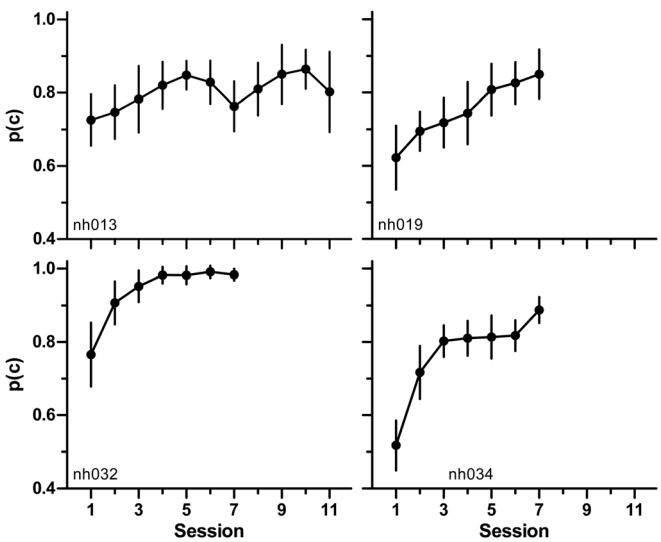

FIG. 5. Same as Fig. 4 for the 4 listeners in Group 2. TF masks for Group 2 were generated with RC = −30 dB.

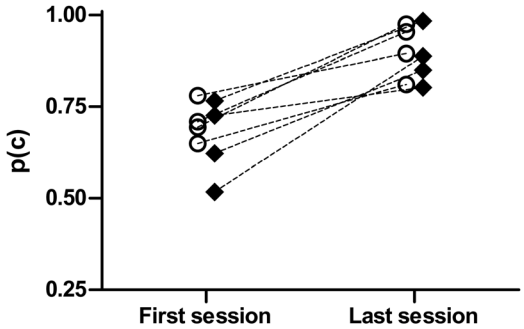

FIG. 6. Change in word recognition from the first to the last session for listeners in the two groups. Each symbol marks the mean score of 1 listener in 1 session. Open symbols: 4 listeners in Group 1. Filled symbols: 4 listeners in Group 2.

The change is shown for the individual listeners in both groups in Fig. 6. The pattern of change was nearly the same for each group with an average improvement in intelligibility of approximately 0.22. A two-way analysis of variance found that the main effect of practice was highly significant ($F^{1,4} = 23.02$, $p < 0.01$). There was no significant difference between groups ($F^{3,4} = 1.72$, $p = 0.30$).

Results from the post-training phase are shown for each group in Fig. 7. Each panel compares pre- and post-training

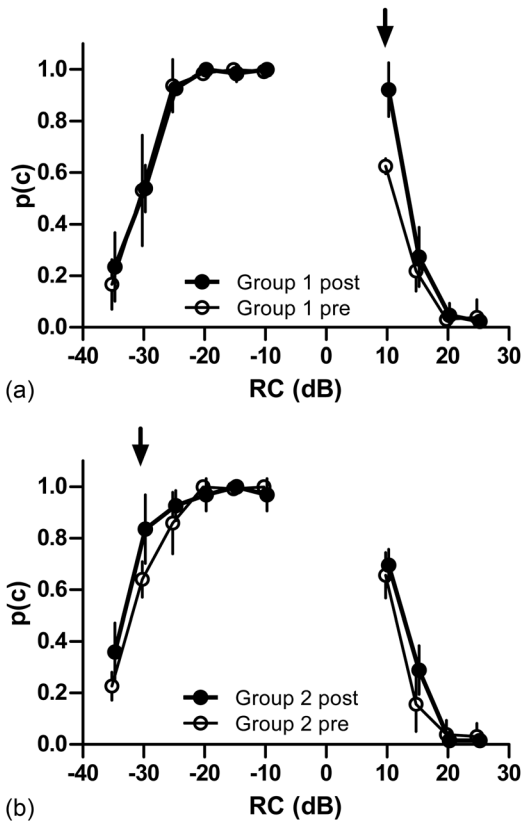

FIG. 7. Post-training word recognition of TF-masked CRM sentences, compared to pre-training scores. (a) Results from Group 1. Group 1 listeners received training with sentences processed by a TF mask using the RC value shown by the arrow (+10 dB). Open symbols: mean word recognition prior to training. Filled symbols: mean word recognition after training. Error bars are standard deviations across listeners. (b) Same as (a) for Group 2. Group 2 listeners received training with sentences processed by a TF mask using the RC value shown by the arrow (−30 dB).

word recognition scores for sentences processed with TF masks generated with a range of values of RC. As expected, each group showed improvement for the RC condition for which training was provided; the magnitude of improvement is similar to that shown in Fig. 6, although the scores shown here were obtained in sessions not included in the previous figure. Recognition scores for the additional RC conditions also tended to improve, but by smaller amounts. This is shown with greater resolution in Fig. 8, which displays the post-training change in recognition score. Forty-two of the sixty-two scores shown in Fig. 8 were greater than zero, indicating improvement after training, and only nine were less than zero. As noted the largest increases were observed at and around the values of RC selected for training, especially for Group 1. The figure also includes nine comparisons of scores obtained for sentences produced by a different male talker, whose sentences were not presented during training. Although the amount of data obtained with the different talker was small, these scores generally fell within the same range as the scores obtained with the same talker whose sentences were presented during training; the clearest examples can be seen for two listeners in Group 1 at RC = 10 and 15 dB. These results suggest that the beneficial effects of training on word recognition did not strongly generalize to sentences processed with TF masks generated with different threshold parameters.

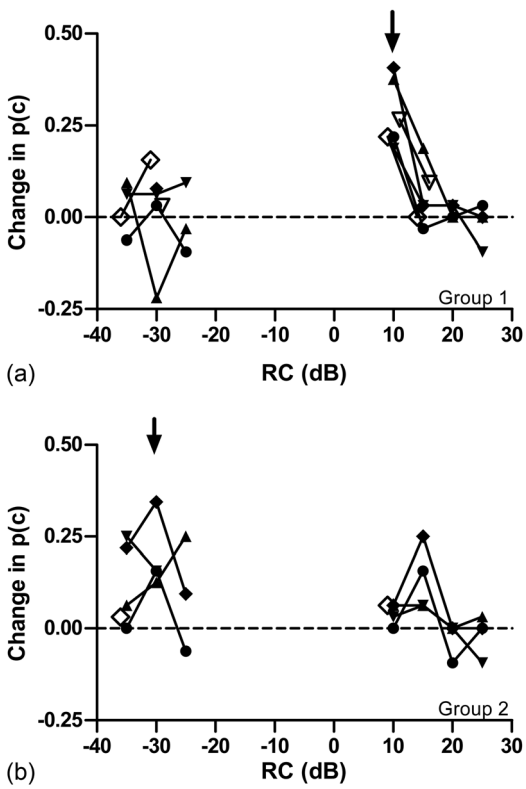

FIG. 8. Post-training change in the recognition of TF-masked CRM sentences. Scores greater than zero indicate improvement. Conditions for which pre-training recognition scores were essentially perfect (RC between −20 and 0 dB) are not shown. Data from individual listeners are shown with different symbols. Filled symbols represent the change in recognition for sentences produced by a single male talker. Open symbols represent the change in recognition for sentences produced by a different male talker, whose sentences were not presented during training. (a) Listeners from Group 1 for which RC during training was +10 dB. (b) Listeners from Group 2 for which RC during training was −30 dB.

IV. DISCUSSION

Application of an ideal TF mask is a powerful method for removing noise from a speech-noise mixture. Wang (2005) has suggested that the separation achieved by an ideal TF mask could serve as a baseline for evaluating other noise rejection methods. With appropriate choice of parameters, ideal TF masks have been shown to restore essentially perfect intelligibility for speech-noise mixtures. However, identification is not always perfect; in particular, previous studies have shown that the intelligibility of speech processed by TF masks is strongly dependent on the criterion used to separate speech from noise.

When recognition scores are plotted as a function of the relative criterion RC, identification is largely independent of overall SNR for SNR as low as −60 dB (Kjems *et al.*, 2009). For any particular SNR, recognition varies with RC in a pattern that can be subdivided into three distinct regions (Brungart *et al.*, 2006; Kjems *et al.*, 2009; Li and Loizou, 2008). For RC values in the range from approximately −20 to 0 dB, identification is essentially perfect. Recognition accuracy decreases outside that range. Pre-training recognition scores reported here were consistent with that pattern, with the qualification that the range of RC values was sampled less densely in this study than in others.

For RC greater than 0 dB, identification scores decrease because the TF mask rejects too much of the speech (Brungart *et al.*, 2006). The TF mask shown in Fig. 2(b) was characterized by a low proportion of cells with gain set to 1, which accounts for the loss of speech information. For RC less than about −20 dB, identification scores also decrease, in that case, because the TF mask rejects too little of the noise masker. The TF mask shown in Fig. 2(c), which was characterized by a high proportion of cells with gain set to 1, illustrates why that happens.

The results presented here showed that extended practice with sentences processed with a particular TF masked produced in each group a consistent increase of about 0.2 in recognition scores. A comparison of pre- and post-training measurements suggested that the improvement was slightly more dramatic for Group 1 [Fig. 8(a)]. However, comparable improvements were not seen for the identification of speech-noise mixtures processed with different TF masks. That may reflect the fact that this study was carried out with a limited stimulus set; additional experiments would be required to determine whether training on a more diverse stimulus set would lead to greater generalization.

The goal of the study reported here was to determine the extent to which practice improves the identification of speech-noise mixtures processed by ideal TF masks generated with threshold criteria that do not lead immediately to perfect identification. This was done because it will eventually be possible to generate non-ideal TF masks that could be used in real-world applications such as hearing aid processors; a non-ideal TF mask is one that is generated without

prior knowledge of the speech signal that the mask must extract. At least initially, non-ideal TF masks will not have the same ability to reject noise as the most-effective ideal TF masks do. Preliminary experiments to generate non-ideal TF masks are underway in this laboratory, and in others (e.g., Kim *et al.*, 2009). The present results suggest that when those TF masks are evaluated, it will be important to provide extended training before concluding that a strategy is ineffective. It cannot be said with certainty whether (or if) sound quality from those TF masks will be degraded in a way that is comparable to the degradation of sound quality produced by the TF masks used in the training conditions in this study. However, two different forms of degradation were examined here: one that rejected too much of the speech (RC = +10 dB) and another that allowed too much noise to pass (RC = −30 dB). Listeners showed improvement in each condition, which may suggest that practice will produce improvement in the intelligibility of speech processed in other as-yet untested ways.

ACKNOWLEDGMENTS

Supported by Grant No. DC010615 from NIDCD to D.G.S. The authors thank Dr. Nandini Iyer for providing a copy of the CRM corpus.

Anzalone, M. C., Calandruccio, L., Doherty, K. A., and Carney, L. H. (**2006**). "Determination of the potential benefit of time-frequency gain manipulation," Ear Hear. **27**, 480–492.

Bolia, R. S., Nelson, T. W., Ericson, M. A., and Simpson, B. (**2000**). "A speech corpus for multitalker communications research," J. Acoust. Soc. Am. **107**, 1065–1066.

Brungart, D. S., Chang, P. S., Simpson, B. D., and Wang, D. L. (**2006**). "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," J. Acoust. Soc. Am. **120**, 4007–4018.

Burk, M. H., and Humes, L. E. (**2007**). "Effects of training on speech-recognition performance in noise using lexically hard words," J. Speech Lang. Hear. Res. **50**, 25–40.

Burk, M. H., and Humes, L. E. (**2009**). "Effects of long-term training on aided speech-recognition performance in noise in older adults," J. Speech Lang. Hear. Res. **51**, 759–771.

Burk, M. H., Humes, L. E., Amos, N., and Strauser, L. (**2006**). "Effect of training on word-recognition performance in noise for young normal-hearing and older hearing-impaired listeners," Ear Hear. **27**, 263–278.

Choi, S., Kirk, K., Talavage, T., Krull, V., Smalt, C., and Baker, S. (**2009**). "Effects of training format on perceptual learning of spectrally degraded voice,s" J. Acoust. Soc. Am. **125**, 2526.

Eddins, D. A., and Li, C. (**2012**). "Psychometric properties of the coordinate response measure corpus with various types of background interference," J. Acoust. Soc. Am. **131**, EL177–EL183.

Helfer, K. S., and Freyman, R. L. (**2008**). "Aging and speech-on-speech masking," Ear Hear. **29**, 87–98.

Kim, G., Lu, Y., Hu, Y., and Loizou, P. C. (**2009**). "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," J. Acoust. Soc. Am. **126**, 1486–1494.

Kjems, U., Boldt, J. B., Pedersen, M. S., Lunner, T., and Wang, D. L. (**2009**). "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," J. Acoust. Soc. Am. **126**, 1415–1426.

Li, N., and Loizou, P. C. (**2008**). "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," J. Acoust. Soc. Am. **123**, 1673–1682.

Moore, B. C. J., and Glasberg, B. R. (**1987**). "Formulae describing frequency selectivity as a function of frequency and level, and their use in calculating excitation patterns," Hear Res. **28**, 209–225.

Narayanan, A., and Wang, D. L. (**2010**). "Robust speech recognition from binary masks," J. Acoust. Soc. Am. **128**, EL217–EL222.

Plomp, R., and Mimpen, A. M. (**1979**). "Speech-reception threshold for sentences as a function of age and noise level," J. Acoust. Soc. Am. **66**, 1333–1342.

Sweetow, R., and Palmer, C. V. (**2005**). "Efficacy of individual auditory training in adults: A systematic review of the evidence," J. Am. Acad. Audiol. **16**, 494–504.

Turner, C. W. (**2006**). "Hearing loss and the limits of amplification," Audiol. Neuro-Otol. **11**, Suppl. 1, 2–5.

Tyler, R. S., Witt, S. A., Dunn, C. C., and Wang, W. (**2010**). "Initial development of a spatially separated speech-in-noise and localization training program," J. Am. Acad. Audiol. **21**, 390–403.

Wang, D. L. (**2005**). "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, edited by P. Divenyi (Kluwer Academic, Norwell, MA), pp. 181–197.

Wang, D. L. (**2008**). "Time-frequency masking for speech separation and its potential for hearing aid design," Trends Amplif. **12**, 332–353.

Wang, D. L., Kjems, U., Pedersen, M. S., Boldt, J. B., and Lunner, T. (**2008**). "Speech perception of noise with binary gains," J. Acoust. Soc. Am. **124**, 2303–2307.

Wang, D. L., Kjems, U., Pedersen, M. S., Boldt, J. B., and Lunner, T. (**2009**). "Speech intelligibility in background noise with ideal binary time-frequency masking," J. Acoust. Soc. Am. **125**, 2336–2347.